

Hypothesis Testing Using Pairwise Distances and Associated Kernels

Dino Sejdinovic*

Arthur Gretton^{*,†,*}

Bharath Sriperumbudur^{*,*}

Kenji Fukumizu[‡]

DINO.SEJDINOVIC@GMAIL.COM

ARTHUR.GRETTON@GMAIL.COM

BHARATH@GATSBY.UCL.AC.UK

FUKUMIZU@ISM.AC.JP

*Gatsby Computational Neuroscience Unit, CSML, University College London, [†]Max Planck Institute for Intelligent Systems, Tübingen, [‡]The Institute of Statistical Mathematics, Tokyo

Abstract

We provide a unifying framework linking two classes of statistics used in two-sample and independence testing: on the one hand, the energy distances and distance covariances from the statistics literature; on the other, distances between embeddings of distributions to reproducing kernel Hilbert spaces (RKHS), as established in machine learning. The equivalence holds when energy distances are computed with semimetrics of negative type, in which case a kernel may be defined such that the RKHS distance between distributions corresponds exactly to the energy distance. We determine the class of probability distributions for which kernels induced by semimetrics are characteristic (that is, for which embeddings of the distributions to an RKHS are injective). Finally, we investigate the performance of this family of kernels in two-sample and independence tests: we show in particular that the energy distance most commonly employed in statistics is just one member of a parametric family of kernels, and that other choices from this family can yield more powerful tests.

1. Introduction

The problem of testing statistical hypotheses in high dimensional spaces is particularly challenging, and has been a recent focus of considerable work in the statistics and machine learning communities. On the statistical side, two-sample testing in Euclidean spaces (of whether two independent samples are from the

same distribution, or from different distributions) can be accomplished using a so-called energy distance as a statistic (Székely & Rizzo, 2004; 2005). Such tests are consistent against all alternatives as long as the random variables have finite first moments. A related dependence measure between vectors of high dimension is the distance covariance (Székely et al., 2007; Székely & Rizzo, 2009), and the resulting test is again consistent for variables with bounded first moment. The distance covariance has had a major impact in the statistics community, with Székely & Rizzo (2009) being accompanied by an editorial introduction and discussion. A particular advantage of energy distance-based statistics is their compact representation in terms of certain expectations of pairwise Euclidean distances, which leads to straightforward empirical estimates. As a follow-up work, Lyons (2011) generalized the notion of distance covariance to metric spaces of negative type (of which Euclidean spaces are a special case).

On the machine learning side, two-sample tests have been formulated based on embeddings of probability distributions into reproducing kernel Hilbert spaces (Gretton et al., 2012), using as the test statistic the difference between these embeddings: this statistic is called the maximum mean discrepancy (MMD). This distance measure was applied to the problem of testing for independence, with the associated test statistic being the Hilbert-Schmidt Independence Criterion (HSIC) (Gretton et al., 2005a; 2008; Smola et al., 2007; Zhang et al., 2011). Both tests are shown to be consistent against all alternatives when a characteristic RKHS is used (Fukumizu et al., 2008; Sriperumbudur et al., 2010). Such tests can further be generalized to structured and non-Euclidean domains, such as text strings, graphs or groups (Fukumizu et al., 2009).

Despite their striking similarity, the link between energy distance-based tests and kernel-based tests

* These authors contributed equally.

Appendix available at arxiv.org/abs/1205.0411.

Appearing in *Proceedings of the 29th International Conference on Machine Learning*, Edinburgh, Scotland, UK, 2012.

Copyright 2012 by the author(s)/owner(s).

has been an open question. In the discussion of Székely & Rizzo (2009), Gretton et al. (2009b, p. 1289) first explored this link in the context of independence testing, and stated that interpreting the distance-based independence statistic as a kernel statistic is not straightforward, since Bochner’s theorem does not apply to the choice of weight function used in the definition of Brownian distance covariance (we briefly review this argument in Section A.3 of the Appendix). Székely & Rizzo (2009, Rejoinder, p. 1303) confirmed this conclusion, and commented that RKHS-based dependence measures do not seem to be formal extensions of Brownian distance covariance because the weight function is not integrable. Our contribution resolves this question and shows that RKHS-based dependence measures are precisely the formal extensions of Brownian distance covariance, where the problem of non-integrability of weight functions is circumvented by using translation-variant kernels, i.e., *distance-induced kernels*, a novel family of kernels that we introduce in Section 2.2.

In the case of two-sample testing, we demonstrate that energy distances are in fact maximum mean discrepancies arising from the same family of distance-induced kernels. A number of interesting consequences arise from this insight: first, we show that the energy distance (and distance covariance) derives from a particular parameter choice from a larger family of kernels: this choice may not yield the most sensitive test. Second, results from Gretton et al. (2009a); Zhang et al. (2011) may be applied to get consistent two-sample and independence tests for the energy distance, without using bootstrap, which perform much better than the upper bound proposed by Székely et al. (2007) as an alternative to the bootstrap. Third, in relation to Lyons (2011), we obtain a new family of characteristic kernels arising from semimetric spaces of negative type (where the triangle inequality need not hold), which are quite unlike the characteristic kernels defined via Bochner’s theorem (Sriperumbudur et al., 2010).

The structure of the paper is as follows: In Section 2, we provide the necessary definitions from RKHS theory, and the relation between RKHS and semimetrics of negative type. In Section 3.1, we review both the energy distance and distance covariance. We relate these quantities in Sections 3.2 and 3.3 to the Maximum Mean Discrepancy (MMD) and the Hilbert-Schmidt Independence Criterion (HSIC), respectively. We give conditions for these quantities to distinguish between probability measures in Section 4, thus obtaining a new family of characteristic kernels. Empirical estimates of these quantities and associated two-sample and independence tests are described in Sec-

tion 5. Finally, in Section 6, we investigate the performance of the test statistics on a variety of testing problems, which demonstrate the strengths of the new kernel family.

2. Definitions and Notation

In this section, we introduce concepts and notation required to understand reproducing kernel Hilbert spaces (Section 2.1), and distribution embeddings into RKHS. We then introduce semimetrics (Section 2.2), and review the relation of semimetrics of negative type to RKHS kernels.

2.1. RKHS Definitions

Unless stated otherwise, we will assume that \mathcal{Z} is any topological space.

Definition 1. (RKHS) Let \mathcal{H} be a Hilbert space of real-valued functions defined on \mathcal{Z} . A function $k : \mathcal{Z} \times \mathcal{Z} \rightarrow \mathbb{R}$ is called a *reproducing kernel* of \mathcal{H} if (i) $\forall z \in \mathcal{Z}, k(\cdot, z) \in \mathcal{H}$, and (ii) $\forall z \in \mathcal{Z}, \forall f \in \mathcal{H}, \langle f, k(\cdot, z) \rangle_{\mathcal{H}} = f(z)$. If \mathcal{H} has a reproducing kernel, it is called a *reproducing kernel Hilbert space* (RKHS).

According to the Moore-Aronszajn theorem (Berlinet & Thomas-Agnan, 2004, p. 19), for every symmetric, positive definite function $k : \mathcal{Z} \times \mathcal{Z} \rightarrow \mathbb{R}$, there is an associated RKHS \mathcal{H}_k of real-valued functions on \mathcal{Z} with reproducing kernel k . The map $\varphi : \mathcal{Z} \rightarrow \mathcal{H}_k, \varphi : z \mapsto k(\cdot, z)$ is called the canonical feature map or the Aronszajn map of k . We will say that k is a nondegenerate kernel if its Aronszajn map is injective.

2.2. Semimetrics of Negative Type

We will work with the notion of semimetric of negative type on a non-empty set \mathcal{Z} , where the “distance” function need not satisfy the triangle inequality. Note that this notion of semimetric is different to that which arises from the seminorm, where distance between two distinct points can be zero (also called pseudonorm).

Definition 2. (Semimetric) Let \mathcal{Z} be a non-empty set and let $\rho : \mathcal{Z} \times \mathcal{Z} \rightarrow [0, \infty)$ be a function such that $\forall z, z' \in \mathcal{Z}$, (i) $\rho(z, z') = 0$ if and only if $z = z'$, and (ii) $\rho(z, z') = \rho(z', z)$. Then (\mathcal{Z}, ρ) is said to be a semimetric space and ρ is called a semimetric on \mathcal{Z} . If, in addition, (iii) $\forall z, z', z'' \in \mathcal{Z}, \rho(z', z'') \leq \rho(z, z') + \rho(z, z'')$, (\mathcal{Z}, ρ) is said to be a metric space and ρ is called a metric on \mathcal{Z} .

Definition 3. (Negative type) The semimetric space (\mathcal{Z}, ρ) is said to have negative type if $\forall n \geq 2$,

$z_1, \dots, z_n \in \mathcal{Z}$, and $\alpha_1, \dots, \alpha_n \in \mathbb{R}$ with $\sum_{i=1}^n \alpha_i = 0$,

$$\sum_{i=1}^n \sum_{j=1}^n \alpha_i \alpha_j \rho(z_i, z_j) \leq 0. \quad (1)$$

Note that in the terminology of Berg et al. (1984), ρ satisfying (1) is said to be a *negative definite* function. The following theorem is a direct consequence of Berg et al. (1984, Proposition 3.2, p. 82).

Proposition 4. *ρ is a semimetric of negative type if and only if there exists a Hilbert space \mathcal{H} and an injective map $\varphi : \mathcal{Z} \rightarrow \mathcal{H}$, such that*

$$\rho(z, z') = \|\varphi(z) - \varphi(z')\|_{\mathcal{H}}^2 \quad (2)$$

This shows that $(\mathbb{R}^d, \|\cdot - \cdot\|^2)$ is of negative type. From Berg et al. (1984, Corollary 2.10, p. 78), we have that:

Proposition 5. *If ρ satisfies (1), then so does ρ^q , for $0 < q < 1$.*

Therefore, by taking $q = 1/2$, we conclude that all Euclidean spaces are of negative type. While Lyons (2011, p. 9) also uses the result in Proposition 4, he studies embeddings to general Hilbert spaces, and the relation with the theory of reproducing kernel Hilbert spaces is not exploited. Semimetrics of negative type and symmetric positive definite kernels are in fact closely related, as summarized in the following Lemma based on Berg et al. (1984, Lemma 2.1, p. 74).

Lemma 6. *Let \mathcal{Z} be a nonempty set, and let ρ be a semimetric on \mathcal{Z} . Let $z_0 \in \mathcal{Z}$, and denote $k(z, z') = \rho(z, z_0) + \rho(z', z_0) - \rho(z, z')$. Then k is positive definite if and only if ρ satisfies (1).*

We call the kernel k defined above the *distance-induced kernel*, and say that it is induced by the semimetric ρ . For brevity, we will drop “induced” hereafter, and say that k is simply the *distance kernel* (with some abuse of terminology). In addition, we will typically work with distance kernels scaled by $1/2$. Note that $k(z_0, z_0) = 0$, so distance kernels are not strictly positive definite (equivalently, $k(\cdot, z_0) = 0$). By varying “the point at the center” z_0 , one obtains a family $\mathcal{K}_\rho = \left\{ \frac{1}{2} [\rho(z, z_0) + \rho(z', z_0) - \rho(z, z')] \right\}_{z_0 \in \mathcal{Z}}$ of distance kernels induced by ρ . We may now express (2) from Proposition 4 in terms of the canonical feature map for the RKHS \mathcal{H}_k (proof in Appendix A.1).

Proposition 7. *Let (\mathcal{Z}, ρ) be a semimetric space of negative type, and $k \in \mathcal{K}_\rho$. Then:*

1. *k is nondegenerate, i.e., the Aronszajn map $z \mapsto k(\cdot, z)$ is injective.*

$$2. \quad \rho(z, z') = k(z, z) + k(z', z') - 2k(z, z') = \|k(\cdot, z) - k(\cdot, z')\|_{\mathcal{H}_k}^2.$$

Note that Proposition 7 implies that the Aronszajn map $z \mapsto k(\cdot, z)$ is an isometric embedding of a metric space $(\mathcal{Z}, \rho^{1/2})$ into \mathcal{H}_k , for every $k \in \mathcal{K}_\rho$.

2.3. Kernels Inducing Semimetrics

We now further develop the link between semimetrics of negative type and kernels. Let k be any nondegenerate reproducing kernel on \mathcal{Z} (for example, every strictly positive definite k is nondegenerate). Then, by Proposition 4,

$$\rho(z, z') = k(z, z) + k(z', z') - 2k(z, z') \quad (3)$$

defines a valid semimetric ρ of negative type on \mathcal{Z} . We will say that k generates ρ . It is clear that every distance kernel $\tilde{k} \in \mathcal{K}_\rho$ also generates ρ , and that \tilde{k} can be expressed as:

$$\tilde{k}(z, z') = k(z, z') + k(z_0, z_0) - k(z, z_0) - k(z', z_0), \quad (4)$$

for some $z_0 \in \mathcal{Z}$. In addition, $k \in \mathcal{K}_\rho$ if and only if $k(z_0, z_0) = 0$ for some $z_0 \in \mathcal{Z}$. Hence, it is clear that any strictly positive definite kernel, e.g., the Gaussian kernel $e^{-\sigma \|z - z'\|^2}$, is *not* a distance kernel.

Example 8. Let $\mathcal{Z} = \mathbb{R}^d$ and write $\rho_q(z, z') = \|z - z'\|^q$. By combining Propositions 4 and 5, ρ_q is a valid semimetric of negative type for $0 < q \leq 2$. It is a metric of negative type if $q \leq 1$. The corresponding distance kernel “centered at zero” is given by

$$k_q(z, z') = \frac{1}{2} (\|z\|^q + \|z'\|^q - \|z - z'\|^q). \quad (5)$$

Example 9. Let $\mathcal{Z} = \mathbb{R}^d$, and consider the Gaussian kernel $k(z, z') = e^{-\sigma \|z - z'\|^2}$. The induced semimetric is $\rho(z, z') = 2 \left[1 - e^{-\sigma \|z - z'\|^2} \right]$. There are many other kernels that generate ρ , however; for example, the distance kernel induced by ρ and “centered at zero” is $\tilde{k}(z, z') = e^{-\sigma \|z - z'\|^2} + 1 - e^{-\sigma \|z\|^2} - e^{-\sigma \|z'\|^2}$.

3. Distances and Covariances

In this section, we begin with a description of the energy distance, which measures distance between distributions; and distance covariance, which measures dependence. We then demonstrate that the former is a special instance of the maximum mean discrepancy (a kernel measure of distance on distributions), and the latter an instance of the Hilbert-Schmidt Independence criterion (a kernel dependence measure). We will denote by $\mathcal{M}(\mathcal{Z})$ the set of all finite signed Borel measures on \mathcal{Z} , and by $\mathcal{M}_+^1(\mathcal{Z})$ the set of all Borel probability measures on \mathcal{Z} .

3.1. Energy Distance and Distance Covariance

Székely & Rizzo (2004; 2005) use the following measure of statistical distance between two probability measures P and Q on \mathbb{R}^d , termed the *energy distance*:

$$D_E(P, Q) = 2\mathbb{E}_{ZW} \|Z - W\| - \mathbb{E}_{ZZ'} \|Z - Z'\| - \mathbb{E}_{WW'} \|W - W'\|, \quad (6)$$

where $Z, Z' \stackrel{i.i.d.}{\sim} P$ and $W, W' \stackrel{i.i.d.}{\sim} Q$. This quantity characterizes the equality of distributions, and in the scalar case, it coincides with twice the Cramer-Von Mises distance. We may generalize it to a semimetric space of negative type (\mathcal{Z}, ρ) , with the expression for this generalized distance covariance $D_{E,\rho}(P, Q)$ being of the same form as (6), with the Euclidean distance replaced by ρ . Note that the negative type of ρ implies the non-negativity of $D_{E,\rho}$. In Section 3.2, we will show that for every ρ , $D_{E,\rho}$ is precisely the MMD associated to a particular kernel k on \mathcal{Z} .

Now, let X be a random vector on \mathbb{R}^p and Y a random vector on \mathbb{R}^q . The distance covariance was introduced in Székely et al. (2007); Székely & Rizzo (2009) to address the problem of testing and measuring dependence between X and Y , in terms of a weighted L_2 -distance between characteristic functions of the joint distribution of X and Y and the product of their marginals. Given a particular choice of weight function, it can be computed in terms of certain expectations of pairwise Euclidean distances,

$$\begin{aligned} \mathcal{V}^2(X, Y) = & \mathbb{E}_{XY} \mathbb{E}_{X'Y'} \|X - X'\| \|Y - Y'\| \\ & + \mathbb{E}_X \mathbb{E}_{X'} \|X - X'\| \mathbb{E}_Y \mathbb{E}_{Y'} \|Y - Y'\| \\ & - 2\mathbb{E}_{X'Y'} [\mathbb{E}_X \|X - X'\| \mathbb{E}_Y \|Y - Y'\|], \end{aligned} \quad (7)$$

where (X, Y) and (X', Y') are $i.i.d.$ P_{XY} . Recently, Lyons (2011) established that the generalization of the distance covariance is possible to metric spaces of negative type, with the expression for this generalized distance covariance $\mathcal{V}_{\rho_X, \rho_Y}^2(X, Y)$ being of the same form as (7), with Euclidean distances replaced by metrics of negative type ρ_X and ρ_Y on domains X and Y , respectively. In Section 3.3, we will show that the generalized distance covariance of a pair of random variables X and Y is precisely HSIC associated to a particular kernel k on the product of domains of X and Y .

3.2. Maximum Mean Discrepancy

The notion of the feature map in an RKHS (Section 2.1) can be extended to kernel embeddings of probability measures (Berlinet & Thomas-Agnan, 2004; Sriperumbudur et al., 2010).

Definition 10. (Kernel embedding) Let k be a kernel on \mathcal{Z} , and $P \in \mathcal{M}_+^1(\mathcal{Z})$. The *kernel embedding*

of P into the RKHS \mathcal{H}_k is $\mu_k(P) \in \mathcal{H}_k$ such that $\mathbb{E}_{Z \sim P} f(Z) = \langle f, \mu_k(P) \rangle_{\mathcal{H}_k}$ for all $f \in \mathcal{H}_k$.

Alternatively, the kernel embedding can be defined by the Bochner expectation $\mu_k(P) = \mathbb{E}_{Z \sim P} k(\cdot, Z)$. By the Riesz representation theorem, a sufficient condition for the existence of $\mu_k(P)$ is that k is Borel-measurable and that $\mathbb{E}_{Z \sim P} k^{1/2}(Z, Z) < \infty$. If k is a bounded continuous function, this is obviously true for all $P \in \mathcal{M}_+^1(\mathcal{Z})$. Kernel embeddings can be used to induce metrics on the spaces of probability measures, giving the maximum mean discrepancy (MMD),

$$\begin{aligned} \gamma_k^2(P, Q) &= \|\mu_k(P) - \mu_k(Q)\|_{\mathcal{H}_k}^2 \\ &= \mathbb{E}_{ZZ'} k(Z, Z') + \mathbb{E}_{WW'} k(W, W') \\ &\quad - 2\mathbb{E}_{ZW} k(Z, W), \end{aligned} \quad (8)$$

where $Z, Z' \stackrel{i.i.d.}{\sim} P$ and $W, W' \stackrel{i.i.d.}{\sim} Q$. If the restriction of μ_k to some $\mathcal{P}(\mathcal{Z}) \subseteq \mathcal{M}_+^1(\mathcal{Z})$ is well defined and injective, then k is said to be characteristic to $\mathcal{P}(\mathcal{Z})$, and it is said to be characteristic (without further qualification) if it is characteristic to $\mathcal{M}_+^1(\mathcal{Z})$. When k is characteristic, γ_k is a metric on $\mathcal{M}_+^1(\mathcal{Z})$, i.e., $\gamma_k(P, Q) = 0$ iff $P = Q$, $\forall P, Q \in \mathcal{M}_+^1(\mathcal{Z})$. Conditions under which kernels are characteristic have been studied by Sriperumbudur et al. (2008); Fukumizu et al. (2009); Sriperumbudur et al. (2010). An alternative interpretation of (8) is as an integral probability metric (Müller, 1997): see Gretton et al. (2012) for details.

In general, distance kernels are continuous but unbounded functions. Thus, kernel embeddings are not defined for all Borel probability measures, and one needs to restrict the attention to a class of Borel probability measures for which $\mathbb{E}_{Z \sim P} k^{1/2}(Z, Z) < \infty$ when discussing the maximum mean discrepancy. We will assume that all Borel probability measures considered satisfy a stronger condition that $\mathbb{E}_{Z \sim P} k(Z, Z) < \infty$ (this reflects a finite first moment condition on random variables considered in distance covariance tests, and will imply that all quantities appearing in our results are well defined). For more details, see Section A.4 in the Appendix. As an alternative to requiring this condition, one may assume that the underlying semimetric space (\mathcal{Z}, ρ) of negative type is itself bounded, i.e., that $\sup_{z, z' \in \mathcal{Z}} \rho(z, z') < \infty$.

We are now able to describe the relation between the maximum mean discrepancy and the energy distance. The following theorem is a consequence of Lemma 6, and is proved in Section A.1 of the Appendix.

Theorem 11. Let (\mathcal{Z}, ρ) be a semimetric space of negative type and let $z_0 \in \mathcal{Z}$. The distance kernel k induced by ρ satisfies $\gamma_k^2(P, Q) = \frac{1}{2} D_{E,\rho}(P, Q)$. In particular, γ_k does not depend on the choice of z_0 .

There is a subtlety to the link between kernels and semimetrics, when used in computing the distance on probabilities. Consider again the family of distance kernels \mathcal{K}_ρ , where the semimetric ρ is itself generated from k according to (3). As we have seen, it may be that $k \notin \mathcal{K}_\rho$, however it is clear that $\gamma_k^2(P, Q) = \frac{1}{2}D_{E,\rho}(P, Q)$ whenever k generates ρ . Thus, all kernels that generate the same semimetric ρ on \mathcal{Z} give rise to the same metric γ_k on (possibly a subset of) $\mathcal{M}_+^1(\mathcal{Z})$, and γ_k is merely an extension of the metric $\rho^{1/2}$ on the point masses. The kernel-based and distance-based methods are therefore equivalent, provided that we allow “distances” ρ which may not satisfy the triangle inequality.

3.3. The Hilbert-Schmidt Independence Criterion

Given a pair of jointly observed random variables (X, Y) with values in $\mathcal{X} \times \mathcal{Y}$, the Hilbert-Schmidt Independence Criterion (HSIC) is computed as the maximum mean discrepancy between the joint distribution P_{XY} and the product of its marginals $P_X P_Y$. Let k_X and k_Y be kernels on \mathcal{X} and \mathcal{Y} , with respective RKHSs \mathcal{H}_{k_X} and \mathcal{H}_{k_Y} . Following Smola et al. (2007, Section 2.3), we consider the MMD associated to the kernel $k((x, y), (x', y')) = k_X(x, x')k_Y(y, y')$ on $\mathcal{X} \times \mathcal{Y}$ with RKHS \mathcal{H}_k isometrically isomorphic to the tensor product $\mathcal{H}_{k_X} \otimes \mathcal{H}_{k_Y}$. It follows that $\theta := \gamma_k^2(P_{XY}, P_X P_Y)$ with

$$\begin{aligned} \theta &= \left\| \mathbb{E}_{XY} [k_X(\cdot, X) \otimes k_Y(\cdot, Y)] \right. \\ &\quad \left. - \mathbb{E}_X k_X(\cdot, X) \otimes \mathbb{E}_Y k_Y(\cdot, Y) \right\|_{\mathcal{H}_{k_X} \otimes \mathcal{H}_{k_Y}}^2 \\ &= \mathbb{E}_{XY} \mathbb{E}_{X'Y'} k_X(X, X') k_Y(Y, Y') \\ &\quad + \mathbb{E}_X \mathbb{E}_{X'} k_X(X, X') \mathbb{E}_Y \mathbb{E}_{Y'} k_Y(Y, Y') \\ &\quad - 2 \mathbb{E}_{X'Y'} [\mathbb{E}_X k_X(X, X') \mathbb{E}_Y k_Y(Y, Y')], \end{aligned}$$

where in the last step we used that $\langle f \otimes g, f' \otimes g' \rangle_{\mathcal{H}_{k_X} \otimes \mathcal{H}_{k_Y}} = \langle f, f' \rangle_{\mathcal{H}_{k_X}} \langle g, g' \rangle_{\mathcal{H}_{k_Y}}$. It can be shown that this quantity is the squared Hilbert-Schmidt norm of the covariance operator between RKHSs (Gretton et al., 2005b). The following theorem demonstrates the link between HSIC and the distance covariance, and is proved in Appendix A.1.

Theorem 12. *Let (\mathcal{X}, ρ_X) and (\mathcal{Y}, ρ_Y) be semimetric spaces of negative type, and $(x_0, y_0) \in \mathcal{X} \times \mathcal{Y}$. Define*

$$\begin{aligned} k((x, y), (x', y')) &:= [\rho_X(x, x_0) + \rho_X(x', x_0) - \rho_X(x, x')] \times \\ &\quad [\rho_Y(y, y_0) + \rho_Y(y', y_0) - \rho_Y(y, y')]. \end{aligned} \quad (9)$$

Then, k is a positive definite kernel on $\mathcal{X} \times \mathcal{Y}$, and $\gamma_k^2(P_{XY}, P_X P_Y) = \mathcal{V}_{\rho_X, \rho_Y}^2(X, Y)$.

We remark that a similar result to Theorem 12 is given by Lyons (2011, Proposition 3.16), but without making use of the RKHS equivalence. Theorem 12 is a more general statement, in the sense that we allow ρ to be a semimetric of negative type, rather than a metric. In addition to yielding a more general statement, the RKHS equivalence leads to a significantly simpler proof: the result is an immediate application of the HSIC expansion of Smola et al. (2007).

4. Distinguishing Probability Distributions

Lyons (2011, Theorem 3.20) shows that distance covariance in a metric space characterizes independence if the metrics satisfy an additional property, termed *strong negative type*. We will extend this notion to a semimetric ρ . We will say that $P \in \mathcal{M}_+^1(\mathcal{Z})$ has a finite first moment w.r.t. ρ if $\int \rho(z, z_0) dP$ is finite for some $z_0 \in \mathcal{Z}$. It is easy to see that the integral $\int \rho d([P - Q] \times [P - Q]) = -D_{E,\rho}(P, Q)$ converges whenever P and Q have finite first moments w.r.t. ρ . In Appendix A.4, we show that this condition is equivalent to $\mathbb{E}_{Z \sim P} k(Z, Z) < \infty$, for a kernel k that generates ρ , which implies the kernel embedding $\mu_k(P)$ is also well defined.

Definition 13. The semimetric space (\mathcal{Z}, ρ) is said to have a *strong negative type* if $\forall P, Q \in \mathcal{M}_+^1(\mathcal{Z})$ with finite first moment w.r.t. ρ ,

$$P \neq Q \Rightarrow \int \rho d([P - Q] \times [P - Q]) < 0. \quad (10)$$

The quantity in (10) is exactly $-2\gamma_k^2(P, Q)$ for all P, Q with finite first moment w.r.t. ρ . We directly obtain:

Proposition 14. *Let kernel k generate ρ . Then (\mathcal{Z}, ρ) has a strong negative type if and only if k is characteristic to all probability measures with finite first moment w.r.t. ρ .*

Thus, the problems of checking whether a semimetric is of strong negative type and whether its associated kernel is characteristic to an appropriate space of Borel probability measures are equivalent. This conclusion has some overlap with Lyons (2011): in particular, Proposition 14 is stated in Lyons (2011, Proposition 3.10), where the barycenter map β is a kernel embedding in our terminology, although Lyons does not consider distribution embeddings in an RKHS.

5. Empirical Estimates and Hypothesis Tests

In the case of two-sample testing, we are given i.i.d. samples $\mathbf{z} = \{z_i\}_{i=1}^m \sim P$ and $\mathbf{w} = \{w_i\}_{i=1}^n \sim Q$. The

empirical (biased) V-statistic estimate of (8) is

$$\hat{\gamma}_{k,V}^2(\mathbf{z}, \mathbf{w}) = \frac{1}{m^2} \sum_{i=1}^m \sum_{j=1}^m k(z_i, z_j) + \frac{1}{n^2} \sum_{i=1}^n \sum_{j=1}^n k(w_i, w_j) - \frac{2}{mn} \sum_{i=1}^m \sum_{j=1}^n k(z_i, w_j). \quad (11)$$

Recall that if we use a distance kernel k induced by a semimetric ρ , this estimate involves only the pairwise ρ -distances between the sample points.

In the case of independence testing, we are given i.i.d. samples $\mathbf{z} = \{(x_i, y_i)\}_{i=1}^m \sim P_{XY}$, and the resulting V-statistic estimate (HSIC) is (Gretton et al., 2005a; 2008)

$$HSIC(\mathbf{z}; k_X, k_Y) = \frac{1}{m^2} \text{Tr}(K_X H K_Y H), \quad (12)$$

where K_X , K_Y and H are $m \times m$ matrices given by $(K_X)_{ij} := k_X(x_i, x_j)$, $(K_Y)_{ij} := k_Y(y_i, y_j)$ and $H_{ij} = \delta_{ij} - \frac{1}{m}$ (centering matrix). As in the two-sample case, if both k_X and k_Y are distance kernels, the test statistic involves only the pairwise distances between the samples, i.e., kernel matrices in (12) may be replaced by distance matrices.

We would like to design distance-based tests with an asymptotic Type I error of α , and thus we require an estimate of the $(1 - \alpha)$ -quantile of the V-statistic distribution under the null hypothesis. Under the null hypothesis, both (11) and (12) converge to a particular weighted sum of chi-squared distributed independent random variables (for more details, see Section A.2). We investigate two approaches, both of which yield consistent tests: a bootstrap approach (Arcones & Giné, 1992), and a spectral approach (Gretton et al., 2009a; Zhang et al., 2011). The latter requires empirical computation of the spectrum of kernel integral operators, a problem studied extensively in the context of kernel PCA (Schölkopf et al., 1997). In the two-sample case, one computes the eigenvalues of the centred Gram matrix $\tilde{K} = H K H$ on the aggregated samples. Here, K is a $2m \times 2m$ matrix, with entries $K_{ij} = k(u_i, u_j)$, $\mathbf{u} = [\mathbf{z} \ \mathbf{w}]$ is the concatenation of the two samples and H is the centering matrix. Gretton et al. (2009a) show that the null distribution defined using these finite sample estimates converges to the population distribution, provided that the spectrum is square-root summable. The same approach can be used for a consistent finite sample null distribution of HSIC, via computation of the eigenvalues of $\tilde{K}_X = H K_X H$ and $\tilde{K}_Y = H K_Y H$ (Zhang et al., 2011).

Both Székely & Rizzo (2004, p. 14) and Székely et al. (2007, p. 2782–2783) establish that the energy distance

and distance covariance statistics, respectively, converge to a particular weighted sum of chi-squares of form similar to that found for the kernel-based statistics. Analogous results for the generalized distance covariance are presented by Lyons (2011, p. 7–8). These works do not propose test designs that attempt to estimate the coefficients in such representations of the null distribution, however (note also that these coefficients have a more intuitive interpretation using kernels). Besides the bootstrap, Székely et al. (2007, Theorem 6) also proposes an independence test using a bound applicable to a general quadratic form Q of centered Gaussian random variables with $\mathbb{E}[Q] = 1$: $\mathbb{P}\{Q \geq (\Phi^{-1}(1 - \alpha/2))^2\} \leq \alpha$, valid for $0 < \alpha \leq 0.215$. When applied to the distance covariance statistic, the upper bound of α is achieved if X and Y are independent Bernoulli variables. The authors remark that the resulting criterion might be over-conservative. Thus, more sensitive tests are possible by computing the spectrum of the centred Gram matrices associated to distance kernels, and we pursue this approach in the next section.

6. Experiments

6.1. Two-sample Experiments

In the two-sample experiments, we investigate three different kinds of synthetic data. In the first, we compare two multivariate Gaussians, where the means differ in one dimension only, and all variances are equal. In the second, we again compare two multivariate Gaussians, but this time with identical means in all dimensions, and variance that differs in a single dimension. In our third experiment, we use the benchmark data of Sriperumbudur et al. (2009): one distribution is a univariate Gaussian, and the second is a univariate Gaussian with a sinusoidal perturbation of increasing frequency (where higher frequencies correspond to harder problems). All tests use a distance kernel induced by the Euclidean distance. As shown on the left plots in Figure 1, the spectral and bootstrap test designs appear indistinguishable, and they significantly outperform the test designed using the quadratic form bound, which appears to be far too conservative for the data sets considered. This is confirmed by checking the Type I error of the quadratic form test, which is significantly smaller than the test size of $\alpha = 0.05$.

We also compare the performance to that of the Gaussian kernel, with the bandwidth set to the median distance between points in the aggregation of samples. We see that when the means differ, both tests perform similarly. When the variances differ, it is clear that the Gaussian kernel has a major advantage over the dis-

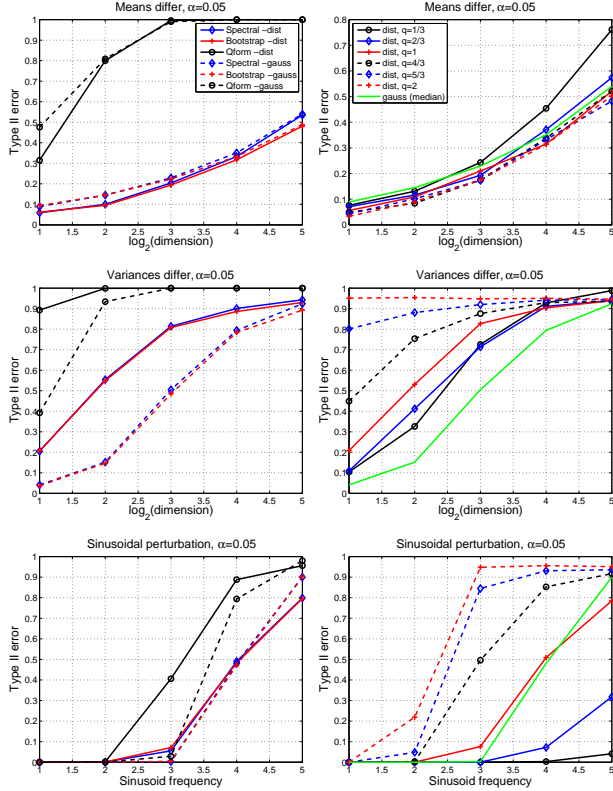


Figure 1. (left) MMD using Gaussian and distance kernels for various tests; (right) Spectral MMD using distance kernels with various exponents. The number of samples in all experiments was set to $m = 200$.

tance kernel, although this advantage decreases with increasing dimension (where both perform poorly). In the case of a sinusoidal perturbation, the performance is again very similar.

In addition, following Example 8, we investigate the performance of kernels obtained using the semimetric $\rho(z, z') = \|z - z'\|^q$ for $0 < q \leq 2$. Results are presented in the right hand plots of Figure 1. While judiciously chosen values of q offer some improvement in the cases of differing mean and variance, we see a dramatic improvement for the sinusoidal perturbation, compared with the case $q = 1$ and the Gaussian kernel: values $q = 1/3$ (and smaller) yield virtually error-free performance even at high frequencies (note that $q = 1$ corresponds to the energy distance described in Székely & Rizzo (2004; 2005)). Additional experiments with real-world data are presented in Appendix A.6.

We observe from the simulation results that distance kernels with higher exponents are advantageous in cases where distributions differ in mean value along a single dimension (with noise in the remainder), whereas distance kernels with smaller exponents are

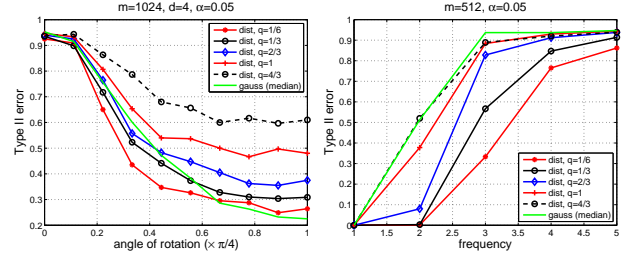


Figure 2. HSIC using distance kernels with various exponents and a Gaussian kernel as a function of (left) the angle of rotation for the dependence induced by rotation; (right) frequency ℓ in the sinusoidal dependence example.

more sensitive to differences in distributions at finer lengthscales (i.e., where the characteristic functions of the distributions differ at higher frequencies). This observation also appears to hold true on the real-world data experiments in Appendix A.6.

6.2. Independence Experiments

To assess independence tests, we used an artificial benchmark proposed by Gretton et al. (2008): we generate univariate random variables from the ICA benchmark densities of Bach & Jordan (2002); rotate them in the product space by an angle between 0 and $\pi/4$ to introduce dependence; fill additional dimensions with independent Gaussian noise; and, finally, pass the resulting multivariate data through random and independent orthogonal transformations. The resulting random variables X and Y are dependent but uncorrelated. The case $m = 1024$ (sample size) and $d = 4$ (dimension) is plotted in Figure 2 (left). As observed by Gretton et al. (2009b), the Gaussian kernel does better than the distance kernel with $q = 1$. By varying q , however, we are able to obtain a wide range of performance; in particular, the values $q = 1/6$ (and smaller) have an advantage over the Gaussian kernel on this dataset, especially in the case of smaller angles of rotation. As for the two-sample case, bootstrap and spectral tests have indistinguishable performance, and are significantly more sensitive than the quadratic form based test, which failed to reject the null hypothesis of independence on this dataset.

In addition, we assess the test performance on sinusoidally dependent data. The distribution over the random variable pair X, Y was drawn from $P_{XY} \propto 1 + \sin(\ell x) \sin(\ell y)$ for integer ℓ , on the support $\mathcal{X} \times \mathcal{Y}$, where $\mathcal{X} := [-\pi, \pi]$ and $\mathcal{Y} := [-\pi, \pi]$. In this way, increasing ℓ caused the departure from a uniform (independent) distribution to occur at increasing frequencies, making this departure harder to detect from a small sample size. Results are in Figure 2 (right). We note that the distance covariance outperforms the

Gaussian kernel on this example, and that smaller exponents result in better performance (lower Type II error when the departure from independence occurs at higher frequencies). Finally, we note that the setting $q = 1$, which is described in Székely et al. (2007); Székely & Rizzo (2009), is a reasonable heuristic in practice, but does not yield the most powerful tests on either dataset.

7. Conclusion

We have established an equivalence between the energy distance and distance covariance, and RKHS measures of distance between distributions. In particular, energy distances and RKHS distance measures coincide when the kernel is induced by a semimetric of negative type. The associated family of kernels performs well in two-sample and independence testing: interestingly, the parameter choice most commonly used in the statistics literature does not yield the most powerful tests in many settings.

The interpretation of the energy distance and distance covariance in an RKHS setting should be of considerable interest both to statisticians and machine learning researchers, since the associated kernels may be used much more widely: in conditional dependence testing and estimates of the chi-squared distance (Fukumizu et al., 2008), in Bayesian inference (Fukumizu et al., 2011), in mixture density estimation (Sriperumbudur, 2011) and in other machine learning applications. In particular, the link with kernels makes these applications of the energy distance immediate and straightforward. Finally, for problem settings defined most naturally in terms of distances, and where these distances are of negative type, there is an interpretation in terms of reproducing kernels, and the learning machinery from the kernel literature can be brought to bear.

References

- Arcones, M. and Giné, E. On the bootstrap of U and V statistics. *Ann. Stat.*, 20(2):655–674, 1992.
- Bach, F. R. and Jordan, M. I. Kernel independent component analysis. *J. Mach. Learn. Res.*, 3:1–48, 2002.
- Berg, C., Christensen, J. P. R., and Ressel, P. *Harmonic Analysis on Semigroups*. Springer, New York, 1984.
- Berlinet, A. and Thomas-Agnan, C. *Reproducing Kernel Hilbert Spaces in Probability and Statistics*. Kluwer, 2004.
- Fukumizu, K., Gretton, A., Sun, X., and Schoelkopf, B. Kernel measures of conditional dependence. In *NIPS*, 2008.
- Fukumizu, K., Sriperumbudur, B., Gretton, A., and Schoelkopf, B. Characteristic kernels on groups and semigroups. In *NIPS*, 2009.
- Fukumizu, K., Song, L., and Gretton, A. Kernel Bayes’ rule. In *NIPS*, 2011.
- Gretton, A., Bousquet, O., Smola, A. J., and Schölkopf, B. Measuring statistical dependence with Hilbert-Schmidt norms. In *ALT*, 2005a.
- Gretton, A., Herbrich, R., Smola, A., Bousquet, O., and Schölkopf, B. Kernel methods for measuring independence. *J. Mach. Learn. Res.*, 6:2075–2129, 2005b.
- Gretton, A., Fukumizu, K., Teo, C. H., Song, L., Schoelkopf, B., and Smola, A. A kernel statistical test of independence. In *NIPS*, 2008.
- Gretton, A., Fukumizu, K., Harchaoui, Z., and Sriperumbudur, B. A fast, consistent kernel two-sample test. In *NIPS*, 2009a.
- Gretton, A., Fukumizu, K., and Sriperumbudur, B. Discussion of: Brownian distance covariance. *Ann. Appl. Stat.*, 3(4):1285–1294, 2009b.
- Gretton, A., Borgwardt, K., Rasch, M., Schölkopf, B., and Smola, A. A kernel two-sample test. *J. Mach. Learn. Res.*, 13:723–773, 2012.
- Lyons, R. Distance covariance in metric spaces. arXiv:1106.5758, June 2011.
- Müller, A. Integral probability metrics and their generating classes of functions. *Adv. Appl. Probab.*, 29(2):429–443, 1997.
- Schölkopf, B., Smola, A. J., and Müller, K.-R. Kernel principal component analysis. In *ICANN*, 1997.
- Smola, A. J., Gretton, A., Song, L., and Schölkopf, B. A Hilbert space embedding for distributions. In *ALT*, 2007.
- Sriperumbudur, B. Mixture density estimation via Hilbert space embedding of measures. In *IEEE ISIT*, 2011.
- Sriperumbudur, B., Gretton, A., Fukumizu, K., Lanckriet, G., and Schölkopf, B. Injective Hilbert space embeddings of probability measures. In *COLT*, 2008.
- Sriperumbudur, B., Fukumizu, K., Gretton, A., Lanckriet, G., and Schoelkopf, B. Kernel choice and classifiability for RKHS embeddings of probability distributions. In *NIPS*, 2009.
- Sriperumbudur, B., Gretton, A., Fukumizu, K., Lanckriet, G., and Schölkopf, B. Hilbert space embeddings and metrics on probability measures. *J. Mach. Learn. Res.*, 11:1517–1561, 2010.
- Steinwart, I. and Christmann, A. *Support Vector Machines*. Springer, 2008.
- Székely, G. and Rizzo, M. Testing for equal distributions in high dimension. *InterStat*, (5), November 2004.
- Székely, G. and Rizzo, M. A new test for multivariate normality. *J. Multivariate Anal.*, 93:58–80, 2005.
- Székely, G. and Rizzo, M. Brownian distance covariance. *Ann. Appl. Stat.*, 4(3):1233–1303, 2009.
- Székely, G., Rizzo, M., and Bakirov, N. K. Measuring and testing dependence by correlation of distances. *Ann. Stat.*, 35(6):2769–2794, 2007.
- Zhang, K., Peters, J., Janzing, D., and Schoelkopf, B. Kernel-based conditional independence test and application in causal discovery. In *UAI*, 2011.

A. Appendix

A.1. Proofs

Proof. (Proposition 7) If $z, z' \in \mathcal{Z}$ are such that $k(w, z) = k(w, z')$, for all $w \in \mathcal{Z}$, one would also have $\rho(z, z_0) - \rho(z, w) = \rho(z', z_0) - \rho(z', w)$, for all $w \in \mathcal{Z}$. In particular, by inserting $w = z$, and $w = z'$, we obtain $\rho(z, z') = -\rho(z, z') = 0$, i.e., $z = z'$. The second statement follows readily by expressing k in terms of ρ . \square

Proof. (Theorem 11) Follows directly by inserting the distance kernel from Lemma 6 into (8), and cancelling out the terms dependant on a single random variable. Define $\theta := \gamma_k^2(P, Q)$.

$$\begin{aligned} \theta &= \frac{1}{2} \mathbb{E}_{ZZ'} [\rho(Z, z_0) + \rho(Z', z_0) - \rho(Z, Z')] \\ &\quad + \frac{1}{2} \mathbb{E}_{WW'} [\rho(W, z_0) + \rho(W', z_0) - \rho(W, W')] \\ &\quad - \mathbb{E}_{ZW} [\rho(Z, z_0) + \rho(W, z_0) - \rho(Z, W)] \\ &= \mathbb{E}_{ZW} \rho(Z, W) - \frac{\mathbb{E}_{ZZ'} \rho(Z, Z')}{2} - \frac{\mathbb{E}_{WW'} \rho(W, W')}{2}. \end{aligned}$$

\square

Proof. (Theorem 12) First, we note that k is a valid reproducing kernel since $k((x, y), (x', y')) = k_{\mathcal{X}}(x, x')k_{\mathcal{Y}}(y, y')$, where we have taken $k_{\mathcal{X}}(x, x') = \rho_{\mathcal{X}}(x, x_0) + \rho_{\mathcal{X}}(x', x_0) - \rho_{\mathcal{X}}(x, x')$, and $k_{\mathcal{Y}}(y, y') = \rho_{\mathcal{Y}}(y, y_0) + \rho_{\mathcal{Y}}(y', y_0) - \rho_{\mathcal{Y}}(y, y')$, as distance kernels induced by $\rho_{\mathcal{X}}$ and $\rho_{\mathcal{Y}}$, respectively. Indeed, a product of two reproducing kernels is always a valid reproducing kernel on the product space (Steinwart & Christmann, 2008, Lemma 4.6, p. 114). To show equality to distance covariance, we start by expanding $\theta := \gamma_k^2(P_{XY}, P_X P_Y)$,

$$\begin{aligned} \theta &= \overbrace{\mathbb{E}_{XY} \mathbb{E}_{X'Y'} k_{\mathcal{X}}(X, X') k_{\mathcal{Y}}(Y, Y')}^{\theta_1} \\ &\quad + \overbrace{\mathbb{E}_X \mathbb{E}_{X'} k_{\mathcal{X}}(X, X') \mathbb{E}_Y \mathbb{E}_{Y'} k_{\mathcal{Y}}(Y, Y')}^{\theta_2} \\ &\quad - \overbrace{2 \mathbb{E}_{X'Y'} [\mathbb{E}_X k_{\mathcal{X}}(X, X') \mathbb{E}_Y k_{\mathcal{Y}}(Y, Y')]}^{\theta_3}. \end{aligned}$$

Note that

$$\begin{aligned} \theta_1 &= \mathbb{E}_{XY} \mathbb{E}_{X'Y'} \rho_{\mathcal{X}}(X, X') \rho_{\mathcal{Y}}(Y, Y') \\ &\quad + 2 \mathbb{E}_X \rho_{\mathcal{X}}(X, x_0) \mathbb{E}_Y \rho_{\mathcal{Y}}(Y, y_0) \\ &\quad + 2 \mathbb{E}_{XY} \rho_{\mathcal{X}}(X, x_0) \rho_{\mathcal{Y}}(Y, y_0) \\ &\quad - 2 \mathbb{E}_{XY} [\rho_{\mathcal{X}}(X, x_0) \mathbb{E}_{Y'} \rho_{\mathcal{Y}}(Y, Y')] \\ &\quad - 2 \mathbb{E}_{XY} [\rho_{\mathcal{Y}}(Y, y_0) \mathbb{E}_{X'} \rho_{\mathcal{X}}(X, X')], \end{aligned}$$

$$\begin{aligned} \theta_2 &= \mathbb{E}_X \mathbb{E}_{X'} \rho_{\mathcal{X}}(X, X') \mathbb{E}_Y \mathbb{E}_{Y'} \rho_{\mathcal{Y}}(Y, Y') \\ &\quad + 4 \mathbb{E}_X \rho_{\mathcal{X}}(X, x_0) \mathbb{E}_Y \rho_{\mathcal{Y}}(Y, y_0) \\ &\quad - 2 \mathbb{E}_X \rho_{\mathcal{X}}(X, x_0) \mathbb{E}_{Y'} \rho_{\mathcal{Y}}(Y, Y') \\ &\quad - 2 \mathbb{E}_Y \rho_{\mathcal{Y}}(Y, y_0) \mathbb{E}_{X'} \rho_{\mathcal{X}}(X, X'), \end{aligned}$$

and

$$\begin{aligned} \theta_3 &= \mathbb{E}_{X'Y'} [\mathbb{E}_X \rho_{\mathcal{X}}(X, X') \mathbb{E}_Y \rho_{\mathcal{Y}}(Y, Y')] \\ &\quad + 3 \mathbb{E}_X \rho_{\mathcal{X}}(X, x_0) \mathbb{E}_Y \rho_{\mathcal{Y}}(Y, y_0) \\ &\quad + \mathbb{E}_{XY} \rho_{\mathcal{X}}(X, x_0) \rho_{\mathcal{Y}}(Y, y_0) \\ &\quad - \mathbb{E}_{XY} [\rho_{\mathcal{X}}(X, x_0) \mathbb{E}_{Y'} \rho_{\mathcal{Y}}(Y, Y')] \\ &\quad - \mathbb{E}_{XY} [\rho_{\mathcal{Y}}(Y, y_0) \mathbb{E}_{X'} \rho_{\mathcal{X}}(X, X')] \\ &\quad - \mathbb{E}_X \rho_{\mathcal{X}}(X, x_0) \mathbb{E}_Y \mathbb{E}_{Y'} \rho_{\mathcal{Y}}(Y, Y') \\ &\quad - \mathbb{E}_Y \rho_{\mathcal{Y}}(Y, y_0) \mathbb{E}_X \mathbb{E}_{X'} \rho_{\mathcal{X}}(X, X'). \end{aligned}$$

The claim now follows by inserting the resulting expansions and cancelling the appropriate terms. Note that only the leading terms in the expansions remain. \square

Remark 15. It turns out that k is not characteristic to $\mathcal{M}_+^1(\mathcal{X} \times \mathcal{Y})$ — i.e., it cannot distinguish between any two distributions on $\mathcal{X} \times \mathcal{Y}$, even if $k_{\mathcal{X}}$ and $k_{\mathcal{Y}}$ are characteristic. However, since γ_k is equal to the Brownian distance covariance, we know that it can always distinguish between any P_{XY} and its product of marginals $P_X P_Y$ in the Euclidean case. Namely, note that $k((x_0, y), (x_0, y')) = k((x, y_0), (x', y_0)) = 0$ for all $x, x' \in \mathcal{X}$, $y, y' \in \mathcal{Y}$. That means that for every two distinct $P_Y, Q_Y \in \mathcal{M}_+^1(\mathcal{Y})$, one has $\gamma_k^2(\delta_{x_0} P_Y, \delta_{x_0} Q_Y) = 0$. Thus, kernel in (9) characterizes independence but not equality of probability measures on the product space. Informally speaking, the independence testing is an easier problem than homogeneity testing on the product space.

A.2. Spectral Tests

Assume that the null hypothesis holds, i.e., that $P = Q$. For a kernel k and a Borel probability measure P , define a kernel “centred” at P : $\tilde{k}_P(z, z') := k(z, z') + \mathbb{E}_{WW'} k(W, W') - \mathbb{E}_W k(z, W) - \mathbb{E}_W k(z', W)$, with $W, W' \stackrel{i.i.d.}{\sim} P$. Note that as a special case for $P = \delta_{z_0}$ we recover the family of kernels in (4), and that $\mathbb{E}_{ZZ'} \tilde{k}_P(Z, Z') = 0$, i.e., $\mu_{\tilde{k}_P}(P) = 0$. The centred kernel is important in characterizing the null distribution of the V-statistic. To the centred kernel \tilde{k}_P on domain \mathcal{Z} , one associates the integral kernel operator $S_{\tilde{k}_P} : L_P^2(\mathcal{Z}) \rightarrow L_P^2(\mathcal{Z})$ (see Steinwart & Christmann, 2008, p. 126–127), given by:

$$S_{\tilde{k}_P} g(z) = \int_{\mathcal{Z}} \tilde{k}_P(z, w) g(w) dP(w). \quad (13)$$

The following theorem is a special case of Gretton et al. (2012, Theorem 12). For simplicity, we focus on the case where $m = n$.

Theorem 16. *Let $\mathbf{Z} = \{Z_i\}_{i=1}^m$ and $\mathbf{W} = \{W_i\}_{i=1}^m$ be two i.i.d. samples from $P \in \mathcal{M}_+^1(\mathcal{Z})$, and let $S_{k_P}^-$ be a trace class operator. Then*

$$\frac{m}{2} \hat{\gamma}_{k,V}^2(\mathbf{Z}, \mathbf{W}) \rightsquigarrow \sum_{i=1}^{\infty} \lambda_i N_i^2, \quad (14)$$

where $N_i \stackrel{i.i.d.}{\sim} \mathcal{N}(0, 1)$, $i \in \mathbb{N}$, and $\{\lambda_i\}_{i=1}^{\infty}$ are the eigenvalues of the operator $S_{k_P}^-$.

Note that this result requires that the integral kernel operator associated to the underlying probability measure P is a trace class operator, i.e., that $\mathbb{E}_{Z \sim P} k(Z, Z) < \infty$. As before, the sufficient condition for this to hold for all probability measures is that k is a bounded function. In the case of a distance kernel, this is the case if the domain \mathcal{Z} has a bounded diameter with respect to the semimetric ρ , i.e., that $\sup_{z, z' \in \mathcal{Z}} \rho(z, z') < \infty$.

The null distribution of HSIC takes an analogous form to (14) of a weighted sum of chi-squares, but with coefficients corresponding to the products of the eigenvalues of integral operators $S_{k_{P_X}}^-$ and $S_{k_{P_Y}}^-$. The following Theorem is in Zhang et al. (2011, Theorem 4) and gives an asymptotic form for the null distribution of HSIC. See also Lyons (2011, Remark 2.9).

Theorem 17. *Let $\mathbf{Z} = \{(X_i, Y_i)\}_{i=1}^m$ be an i.i.d. sample from $P_{XY} = P_X P_Y$, with values in $\mathcal{X} \times \mathcal{Y}$. Let $S_{k_{P_X}}^- : L_{P_X}^2(\mathcal{X}) \rightarrow L_{P_X}^2(\mathcal{X})$, and $S_{k_{P_Y}}^- : L_{P_Y}^2(\mathcal{Y}) \rightarrow L_{P_Y}^2(\mathcal{Y})$ be trace class operators. Then*

$$m \text{HSIC}(\mathbf{Z}; k_X, k_Y) \rightsquigarrow \sum_{i=1}^{\infty} \sum_{j=1}^{\infty} \lambda_i \eta_j N_{i,j}^2, \quad (15)$$

where $N_{i,j} \sim \mathcal{N}(0, 1)$, $i, j \in \mathbb{N}$, are independent and $\{\lambda_i\}_{i=1}^{\infty}$ and $\{\eta_j\}_{j=1}^{\infty}$ are the eigenvalues of the operators $S_{k_{P_X}}^-$ and $S_{k_{P_Y}}^-$, respectively.

Note that if \mathcal{X} and \mathcal{Y} have bounded diameters w.r.t. ρ_X and ρ_Y , Theorem 17 applies to distance kernels induced by ρ_X and ρ_Y for all $P_X \in \mathcal{M}_+^1(\mathcal{X})$, $P_Y \in \mathcal{M}_+^1(\mathcal{Y})$.

A.3. A Characteristic Function Based Interpretation

The distance covariance in (7) was defined by Székely et al. (2007) in terms of a weighted distance between characteristic functions. We briefly review this interpretation here, however we show that this

approach *cannot* be used to derive a kernel-based measure of dependence (this result was first noted by Gretton et al. (2009b)), and is included here in the interests of completeness). Let X be a random vector on $\mathcal{X} = \mathbb{R}^p$ and Y a random vector on $\mathcal{Y} = \mathbb{R}^q$. The characteristic function of X and Y , respectively, will be denoted by f_X and f_Y , and their joint characteristic function by f_{XY} . The distance covariance $\mathcal{V}(X, Y)$ is defined via the norm of $f_{XY} - f_X f_Y$ in a weighted L_2 space on \mathbb{R}^{p+q} , i.e.,

$$\mathcal{V}^2(X, Y) = \int |f_{X,Y}(t, s) - f_X(t) f_Y(s)|^2 w(t, s) dt ds, \quad (16)$$

for a particular choice of weight function given by

$$w(t, s) = \frac{1}{c_p c_q} \cdot \frac{1}{\|t\|^{1+p} \|s\|^{1+q}}, \quad (17)$$

where $c_d = \pi^{\frac{1+d}{2}} / \Gamma(\frac{1+d}{2})$, $d \geq 1$. An important aspect of distance covariance is that $\mathcal{V}(X, Y) = 0$ if and only if X and Y are independent. We next obtain a similar statistic in the kernel setting. Write $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$, and let $k(z, z') = \kappa(z - z')$ be a translation invariant RKHS kernel on \mathcal{Z} , where $\kappa : \mathcal{Z} \rightarrow \mathbb{R}$ is a bounded continuous function. Using Bochner's theorem, κ can be written as:

$$\kappa(z) = \int e^{-z^\top u} d\Lambda(u),$$

for a finite non-negative Borel measure Λ . It follows Gretton et al. (2009b) that

$$\gamma_k^2(P_{XY}, P_X P_Y) = \int |f_{X,Y}(t, s) - f_X(t) f_Y(s)|^2 d\Lambda(t, s),$$

which is in clear correspondence with (16). However, the weight function in (17) is not integrable — so one cannot find a translation invariant kernel for which γ_k coincides with the distance covariance. By contrast, note the kernel in (9) is *not* translation invariant.

A.4. Restriction on Probability Measures

In general, distance kernels and their products are continuous but unbounded, so kernel embeddings are not defined for all Borel probability measures. Thus, one needs to restrict the attention to a particular class of Borel probability measures for which kernel embeddings exist, and a sufficient condition for this is that $\mathbb{E}_{Z \sim P} k^{1/2}(Z, Z) < \infty$, by the Riesz representation theorem. Let k be a measurable reproducing kernel on \mathcal{Z} , and denote, for $\theta > 0$,

$$\mathcal{M}_k^\theta(\mathcal{Z}) = \left\{ \nu \in \mathcal{M}(\mathcal{Z}) : \int k^\theta(z, z) d|\nu|(z) < \infty \right\}. \quad (18)$$

Note that the maximum mean discrepancy $\gamma_k(P, Q)$ is well defined $\forall P, Q \in \mathcal{M}_k^{1/2}(\mathcal{Z}) \cap \mathcal{M}_+^1(\mathcal{Z})$.

Now, let ρ be a semimetric of negative type. Then, we can consider the class of probability measures that have a finite θ -moment with respect to ρ :

$$\mathcal{M}_\rho^\theta(\mathcal{Z}) = \{\nu \in \mathcal{M}(\mathcal{Z}) : \exists z_0 \in \mathcal{Z}, \quad (19) \\ \text{s.t. } \int \rho^\theta(z, z_0) d|\nu|(z) < \infty\}.$$

To ensure existence of energy distance $D_{E,\rho}(P, Q)$, we need to assume that $P, Q \in \mathcal{M}_\theta^1(\mathcal{Z})$, as otherwise expectations $\mathbb{E}_{ZZ'}\rho(Z, Z')$, $\mathbb{E}_{WW'}\rho(W, W')$ and $\mathbb{E}_{ZW}\rho(Z, W)$ may be undefined. The following proposition shows that the classes of probability measures in (18) and (19) coincide at $\theta = n/2$, for $n \in \mathbb{N}$, whenever ρ is generated by kernel k .

Proposition 18. *Let k be a kernel that generates semimetric ρ , and let $n \in \mathbb{N}$. Then, $\mathcal{M}_k^{n/2}(\mathcal{Z}) = \mathcal{M}_\rho^{n/2}(\mathcal{Z})$. In particular, if k_1 and k_2 generate the same semimetric ρ , then $\mathcal{M}_{k_1}^{n/2}(\mathcal{Z}) = \mathcal{M}_{k_2}^{n/2}(\mathcal{Z})$.*

Proof. Let $\theta \geq \frac{1}{2}$. Note that $a^{2\theta}$ is a convex function of a . Suppose $\nu \in \mathcal{M}_k^\theta(\mathcal{Z})$. Then, we have

$$\begin{aligned} & \int \rho^\theta(z, z_0) d|\nu|(z) \\ &= \int \|k(\cdot, z) - k(\cdot, z_0)\|_{\mathcal{H}_k}^{2\theta} d|\nu|(z) \\ &\leq \int (\|k(\cdot, z)\|_{\mathcal{H}_k} + \|k(\cdot, z_0)\|_{\mathcal{H}_k})^{2\theta} d|\nu|(z) \\ &\leq 2^{2\theta-1} \left(\int \|k(\cdot, z)\|_{\mathcal{H}_k}^{2\theta} d|\nu|(z) + \int \|k(\cdot, z_0)\|_{\mathcal{H}_k}^{2\theta} d|\nu|(z) \right) \\ &= 2^{2\theta-1} \left(\int k^\theta(z, z) d|\nu|(z) + k^\theta(z_0, z_0) |\nu|(\mathcal{Z}) \right) \\ &< \infty, \end{aligned}$$

where we have invoked the Jensen's inequality for convex functions. From the above it is clear that $\mathcal{M}_k^\theta(\mathcal{Z}) \subset \mathcal{M}_\rho^\theta(\mathcal{Z})$, for $\theta \geq 1/2$.

To prove the other direction, we show by induction that $\mathcal{M}_\rho^\theta(\mathcal{Z}) \subset \mathcal{M}_k^{n/2}(\mathcal{Z})$ for $\theta \geq \frac{n}{2}$, $n \in \mathbb{N}$. Let $n = 1$. Let $\theta \geq \frac{1}{2}$, and suppose that $\nu \in \mathcal{M}_\rho^\theta(\mathcal{Z})$. Then, by invoking the reverse triangle and Jensen's inequalities, we have:

$$\begin{aligned} \int \rho^\theta(z, z_0) d|\nu|(z) &= \int \|k(\cdot, z) - k(\cdot, z_0)\|_{\mathcal{H}_k}^{2\theta} d|\nu|(z) \\ &\geq \int \left| k^{1/2}(z, z) - k^{1/2}(z_0, z_0) \right|^{2\theta} d|\nu|(z) \\ &\geq \left| \int k^{1/2}(z, z) d|\nu|(z) - \|\nu\|_{TV} k^{1/2}(z_0, z_0) \right|^{2\theta}, \end{aligned}$$

which implies $\nu \in \mathcal{M}_k^{1/2}(\mathcal{Z})$, thereby satisfying the result for $n = 1$. Suppose the result holds for $\theta \geq \frac{n-1}{2}$, i.e., $\mathcal{M}_\rho^\theta(\mathcal{Z}) \subset \mathcal{M}_k^{(n-1)/2}(\mathcal{Z})$ for $\theta \geq \frac{n-1}{2}$. Let $\nu \in \mathcal{M}_\rho^\theta(\mathcal{Z})$ for $\theta \geq \frac{n}{2}$. Then we have

$$\begin{aligned} & \int \rho^\theta(z, z_0) d|\nu|(z) \\ &= \int (\|k(\cdot, z) - k(\cdot, z_0)\|_{\mathcal{H}_k}^2)^{\frac{\theta}{2}} d|\nu|(z) \\ &\geq \left(\int \|k(\cdot, z) - k(\cdot, z_0)\|_{\mathcal{H}_k}^2 d|\nu|(z) \right)^{\frac{\theta}{2}} \\ &\geq \left(\int \left| \|k(\cdot, z)\|_{\mathcal{H}_k} - \|k(\cdot, z_0)\|_{\mathcal{H}_k} \right|^2 d|\nu|(z) \right)^{\frac{\theta}{2}} \\ &\geq \left| \int (\|k(\cdot, z)\|_{\mathcal{H}_k} - \|k(\cdot, z_0)\|_{\mathcal{H}_k})^2 d|\nu|(z) \right|^{\frac{\theta}{2}} \\ &= \left| \int \sum_{r=0}^n (-1)^r \binom{n}{r} \|k(\cdot, z)\|_{\mathcal{H}_k}^{n-r} \|k(\cdot, z_0)\|_{\mathcal{H}_k}^r d|\nu|(z) \right|^{\frac{\theta}{2}} \\ &= \underbrace{\left| \int k^{\frac{n}{2}}(z, z) d|\nu|(z) \right|^{\frac{\theta}{2}}}_A \\ &\quad + \underbrace{\sum_{r=1}^n (-1)^r \binom{n}{r} k^{\frac{r}{2}}(z_0, z_0) \int k^{\frac{n-r}{2}}(z, z) d|\nu|(z) \right|^{\frac{\theta}{2}}}_B. \end{aligned}$$

Note that the terms in B are finite as for $\theta \geq \frac{n}{2} \geq \frac{n-1}{2} \geq \dots \geq \frac{1}{2}$, we have $\mathcal{M}_\rho^\theta(\mathcal{Z}) \subset \mathcal{M}_k^{(n-1)/2}(\mathcal{Z}) \subset \dots \subset \mathcal{M}_k^1(\mathcal{Z}) \subset \mathcal{M}_k^{1/2}(\mathcal{Z})$ and therefore A is finite, which means $\nu \in \mathcal{M}_k^{n/2}(\mathcal{Z})$, i.e., $\mathcal{M}_\rho^\theta(\mathcal{Z}) \subset \mathcal{M}_k^{n/2}(\mathcal{Z})$ for $\theta \geq \frac{n}{2}$. The result shows that $\mathcal{M}_\rho^\theta(\mathcal{Z}) = \mathcal{M}_k^\theta(\mathcal{Z})$ for all $\theta \in \{\frac{n}{2} : n \in \mathbb{N}\}$. \square

The above Proposition gives a natural interpretation of conditions on probability measures in terms of moments w.r.t. ρ . Namely, the kernel embedding $\mu_k(P)$, where kernel k generates the semimetric ρ , exists for every P with finite half-moment w.r.t. ρ , and thus, MMD between P and Q , $\gamma_k(P, Q)$ is well defined whenever both P and Q have finite half-moments w.r.t. ρ . If, in addition, P and Q have finite first moments w.r.t. ρ , then the ρ -energy distance between P and Q is also well defined and it must be equal to the MMD, by Theorem 11.

Rather than imposing the condition on Borel probability measures, one may assume that the underlying semimetric space (\mathcal{Z}, ρ) of negative type is itself bounded, i.e., that $\sup_{z, z' \in \mathcal{Z}} \rho(z, z') < \infty$, implying that distance kernels are bounded functions, and that both MMD and energy distance are always defined. Conversely, bounded kernels (such as Gaussian) always induce bounded semimetrics.

Table 1. MMD with distance kernels on data from Gretton et al. (2009a). Dimensionality is: *Neural I* (64), *Neural II* (100), *Health status* (12,600), *Subtype* (2,118). The boldface denotes instances where distance kernel had smaller Type II error in comparison to Gaussian kernel.

		Gauss	dist (1/3)	dist (2/3)	dist (1)	dist (4/3)	dist (5/3)	dist (2)
<i>Neural I</i>	1- Type I	.956	.969	.964	.949	.952	.959	.959
(<i>m</i> = 200)	Type II	.118	.170	.139	.119	.109	.089	.117
<i>Neural I</i>	1- Type I	.950	.969	.946	.962	.947	.930	.953
(<i>m</i> = 250)	Type II	.063	.075	.045	.041	.040	.065	.052
<i>Neural II</i>	1- Type I	.956	.968	.965	.963	.956	.958	.943
(<i>m</i> = 200)	Type II	.292	.485	.346	.319	.297	.280	.290
<i>Neural II</i>	1- Type I	.963	.980	.968	.950	.952	.960	.941
(<i>m</i> = 250)	Type II	.195	.323	.197	.189	.194	.169	.183
<i>Subtype</i>	1- Type I	.975	.974	.977	.971	.966	.962	.966
(<i>m</i> = 10)	Type II	.055	.828	.237	.092	.042	.033	.024
<i>Health st.</i>	1- Type I	.958	.980	.953	.940	.954	.954	.955
(<i>m</i> = 20)	Type II	.036	.037	.039	.081	.114	.120	.165

A.5. Distance Correlation

The notion of distance covariance extends naturally to that of *distance variance* $\mathcal{V}^2(X) = \mathcal{V}^2(X, X)$ and that of *distance correlation* (in analogy to the Pearson product-moment correlation coefficient):

$$\mathcal{R}^2(X, Y) = \begin{cases} \frac{\mathcal{V}^2(X, Y)}{\mathcal{V}(X)\mathcal{V}(Y)}, & \mathcal{V}(X)\mathcal{V}(Y) > 0, \\ 0, & \mathcal{V}(X)\mathcal{V}(Y) = 0. \end{cases}$$

Distance correlation also has a straightforward interpretation in terms of kernels as:

$$\begin{aligned} \mathcal{R}^2(X, Y) &= \frac{\mathcal{V}^2(X, Y)}{\mathcal{V}(X)\mathcal{V}(Y)} \\ &= \frac{\gamma_k^2(P_{XY}, P_X P_Y)}{\gamma_k(P_{XX}, P_X P_X) \gamma_k(P_{YY}, P_Y P_Y)} \\ &= \frac{\|\Sigma_{XY}\|_{HS}^2}{\|\Sigma_{XX}\|_{HS} \|\Sigma_{YY}\|_{HS}}, \end{aligned}$$

where covariance operator $\Sigma_{XY} : \mathcal{H}_{k_X} \rightarrow \mathcal{H}_{k_Y}$ is a linear operator for which $\langle \Sigma_{XY} f, g \rangle_{\mathcal{H}_{k_Y}} = \mathbb{E}_{XY}[f(X)g(Y)] - \mathbb{E}_X f(X) \mathbb{E}_Y g(Y)$, for all $f \in \mathcal{H}_{k_X}$ and $g \in \mathcal{H}_{k_Y}$, and $\|\cdot\|_{HS}$ denotes the Hilbert-Schmidt norm (Gretton et al., 2005b). It is clear that \mathcal{R} is invariant to scaling $(X, Y) \mapsto (\epsilon X, \epsilon Y)$, $\epsilon > 0$, whenever the corresponding semimetrics are homogeneous, i.e., whenever $\rho_X(\epsilon x, \epsilon x') = \epsilon \rho_X(x, x')$, and similarly for ρ_Y . Moreover, \mathcal{R} is invariant to translations $(X, Y) \mapsto (X + x', Y + y')$, $x' \in \mathcal{X}$, $y' \in \mathcal{Y}$, whenever ρ_X and ρ_Y are translation invariant.

A.6. Further Experiments

We assessed performance of two-sample tests based on distance kernels with various exponents and compared it to that of a Gaussian kernel on real-world multivariate datasets: *Health st.* (microarray data from normal and tumor tissues), *Subtype* (microarray data from different subtypes of cancer) and *Neural I/II* (local field potential (LFP) electrode recordings from the Macaque primary visual cortex (V1) with and without spike events), all discussed in Gretton et al. (2009a). In contrast to Gretton et al. (2009a), we used smaller sample sizes, so that some Type II error persists. At higher sample sizes, all tests exhibit Type II error which is virtually zero. The results are reported in Table 1 below. We used the spectral test for all experiments, and the reported averages are obtained by running 1000 trials. We note that for dataset *Subtype* which is high dimensional but with only a small number of dimensions varying in mean, a larger exponent results in a test of greater power.